Comparison of Optimization Methods for L1-regularized Logistic Regression

Aleksandar Jovanovich
Department of Computer Science and Information Systems
Youngstown State University
Youngstown, OH 44555
aleksjovanovich@gmail.com

Alina Lazar
Department of Computer Science and Information Systems
Youngstown State University
Youngstown, OH 44555
alazar@ysu.edu

Abstract

Logistic regression with L1-regularization has been recognized as a prominent method for feature extraction in linear classification problems. Various optimization methods for L1 logistic regression have been proposed in recent years. However there have been few studies conducted to compare such methods. This paper reviews existing methods for optimization and then tests the methods over a binary dataset. Results are recorded and comparisons are made. After analyzing the results, the conclusion is that the GLMNET method is the best in terms of time efficiency.

Introduction

Digital information is growing at an extreme rate. Emerging technologies have created an environment that is information driven. From social media to medical records, data is collected in all forms from around the world. Current trends suggest a jump in information gathered and collected over the next decade and beyond. Never before has there been an abundance of data and information as we see today.

As the amount of data collected continues to grow so does the challenge of processing and gathering information. The data is growing wide, and the amount of attributes and features that can be derived sometimes outnumber the sample size. Now, more and more binary large objects are appearing in databases which require a different approach to identifying and extracting information.

Researchers have turned to regularized general linear models to form relationships about the binary data. Regularization is required to avoid over-fitting when there are a large number of parameters. In particular, L1-regularized regression is often used for feature selection, and has been shown to generate sparse models (Yuan, Chang, and Lin 2010).

Recently, there has been a large amount of research conducted to related regularization methods. Each method is differentiated by various aspects including: convergence speed, implementation, and practicability. Therefore, there is significance in conducting a thorough comparison and evaluation (Yuan, Chang, and Lin 2010). In this paper, we review prevailing methods for L1-regularized logistic regression and give a detailed comparison.

Background

Logistic regression is used for prediction of the probability of occurrence of an event by fitting data to a function. It is a generalized linear model used for binomial regression. Like other forms of regression analysis, it makes use of one or more predictor variables that may be either numerical or categorical. The logistic regression problem is an optimization problem, and can be solved by a wide variety of methods; such as gradient descent, steepest descent, and Newton. Once optimization is complete and maximum likelihood values are found, a prediction on the probability of the two possible outcomes can be made (Koh, Kim, and Boyd 2007).

The logistic model has the form:

$$\mathrm{Prob}(b|x) = \frac{\exp(b(w^T x + v))}{1 + \exp(b(w^T x + v))} \qquad [1]$$

Where $b \in$ (-1, +1) denotes the associated binary output and where Prob(b|x) is the conditional probability of b.

L1-regularized logistic regression has recently received attention. The main motivation is that L1-regularized

logistic regression yields a sparse vector and has relatively few nonzero coefficients (Koh et al. 2007). A logistic model with sparse vectors is simpler and more efficient when dealing with data having a smaller number of observations than features. When compared to L2-regularized logistic regression, L1-regularized logistic regression outperforms L2-regularized logistic regression (Wainwright, Ravikumar, and Lafferty 2007).
The L1-regularized logistic regression problem minimizes the following equation:

$$l\text{avg}(v,w)+1\|w\|1=(1=m)\sum_{i=1}^{m} f(w^{T}ai+vbi)+1\sum_{i=1}^{n}\|\text{w}\| \quad [2]$$

Where λ > 0 is the regularization parameter. A solution must exist, but it need not be exclusive. The objective function in the L1-regularized Logistic regression problem is not differentiable so solving the problem is a computational challenge (Koh, Kim, and Boyd 2007).

A regularization path is the set of solutions obtained from L1-regularized linear regression problems while solving for λ. In many cases, the entire regularization path needs to be computed, in order to determine an appropriate value of λ. The regularization path in a smaller L1-regularized linear regression problem can be computed efficiently (Friedman, Hastie, and Tibshirani 2010). Hastie et al. describe an algorithm for computing the entire regularization path for general linear models including logistic regression models. Path-following methods can be slow for large-scale problems, where the number of observations is very large.

Optimization

Each method uses a type of optimization approach to find the regularization path as well as λ. The general model used in each method consists of iterations of the descent, where a chosen subset of variables is deemed the working set and all other variables become fixed. With every step the resulting sub-problem contains fewer variables and therefore solved easier.

Coordinate Descent Method

Typically, a coordinate descent method sequentially goes through all variables and then repeats the same process. By solving the regression problem along an entire path of values, this method efficiently calculates the regularization parameters (Friedman, Hastie, and Tibshirani 2010).

Generalized Linear Model with Elastic Net

GLMNET applies a shrinking technique to solve smaller optimization problems. GLMNET conducts feature-wise normalization before solving the optimization problem. Then, GLMNET measures the relative step change in the successive coordinate descent iterations (Yuan, Chang, and Lin 2010).

Continuous Generalized Gradient Descent

An effective regularization strategy in generalized regression is using validation methods to choose a suitable point in a trajectory or a family. Due to the use of gradient information, the number of iterations is less than cyclic coordinate descent methods. However, the cost per iteration is higher (Zhang 2007).

Least Angle Regression

LARS relates to the classic model-selection method known as Forward Selection (described in Efron, Hastie, Johnstone and Tibshirani 2004). Given a collection of possible predictors, a selection is made based on the largest absolute correlation with the response y. Thereafter simple linear regression is performed on the response y. This leaves a residual vector that can be considered the response. Projection is made over the other predictors orthogonally to the response. The selection process is then repeated. After n steps this results in a set of predictors that are then used to construct a n-parameter linear model.

Relaxed Lasso

Relaxo is a generalization of the Lasso shrinkage technique for linear regression. Both variable selection and parameter estimation is achieved by regular Lasso, yet both steps do not necessarily use the same penalty parameter. The results include all Lasso solutions but allow for sparser models while having similar predictive performance if many predictor variables are present. The package is based on the LARS package (Meinshausen 2007).

Datasets

All the experiments were done using the *Leukemia* dataset, a gene-expression data. This dataset was first mentioned in (Golub et al. 1999). The pre-processed dataset using methods from (Dettling, 2004) was used. The datasets consists of 72 genes that are part of two classes 0 and 1. There are 47 genes are from class 0 and 25 are from class 1.

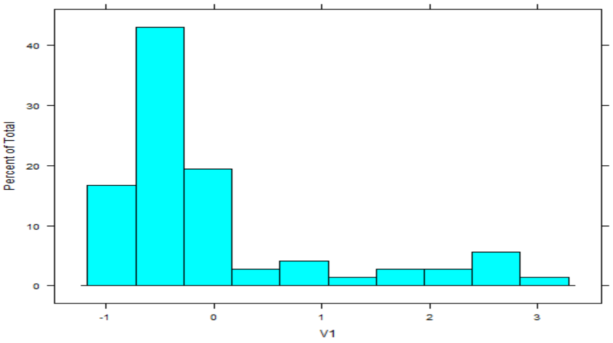

Figure 1. Histogram for Predictor Variable1

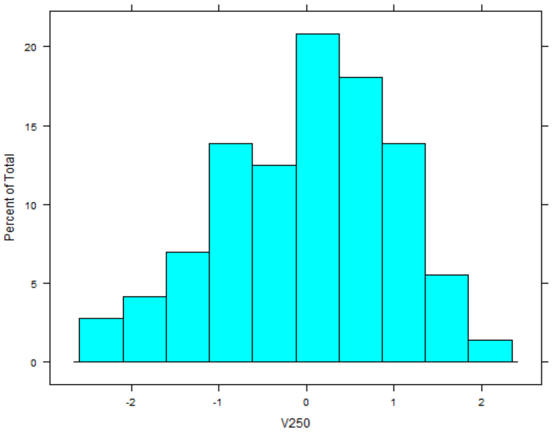

Figure 2. Histogram for Predictor Variable 250

There are 3,571 predictor variables that have numeric values in the interval [-10, 10] with most of the values close to 0.

The two figures above represent the histograms for two of the variables, the first one and the 250th one. More than 75% of the values of variable 1 are in the [-1, 0] interval. The values of variable 250 are normally distributed in the [-2.5, 2.5] interval.

Experiments

So far, we have described several large-scale optimization methods for solving L1-regularized logistic regression problems. In this section, we conduct experiments to investigate their individual and group performances. First we describe the experimental settings. Then the optimization methods are compared in terms of accuracy and time.

To be able to provide good predictions using the GLMNET algorithm, the regularized parameter λ has to be found first. That can be done in R using a grid search and functions from the caret package (Kuhn, 2012). First, the trainControl function is used to set the training parameters. Bootstrap sampling is done 25 times to increase the chance of getting high accuracy results.

```
model <- train(FL,data=trainset,method='glmnet',
                metric = "ROC",
                tuneGrid = expand.grid(.alpha=c(0,1),
                .lambda=seq(0.02,.4,length=20)),
                trControl=MyTrainControl)
```

The model is obtained by using the caret's train function. The search interval for λ is [0.02, .4] with a step of 0.02. Parameter α can take 2 values 0 or 1. For α = 0 and all λ values the AUC (area under the curve) is maximum at 0.992. These results are shown in Figure 3.

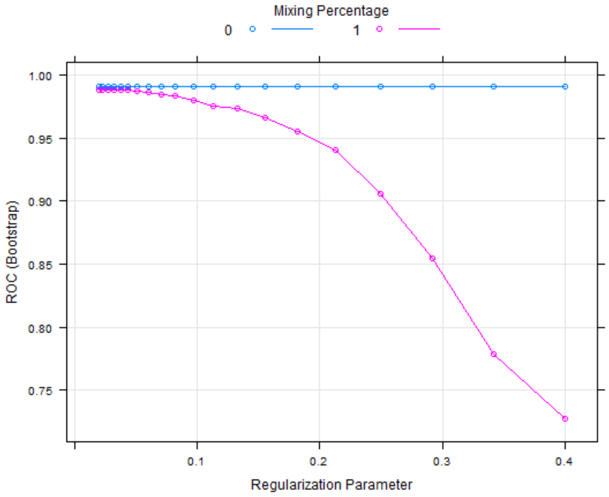

Figure 3.Glmnet ROC curve for the grids search

To run the experiments we used the GLMNET, CGGD, Relaxo, and LARS package in R. The LARS and Relaxo packages fit lasso model paths, while the GLMNET package fits lasso and elastic-net model paths for logistic and multinomial regression using coordinate descent. The algorithms are extremely fast, because they exploit sparsity in the data matrix. The CGGD is used for performing regressions while continuously varying regularization. The method returns the models fit along the continuous paths of parameter modification.

The coefficients from step 1 to 100 were recorded and their profile is plotted in figures 4, 5 and 6. Unfortunately we were unable to plot the coefficients of the Relaxo package.

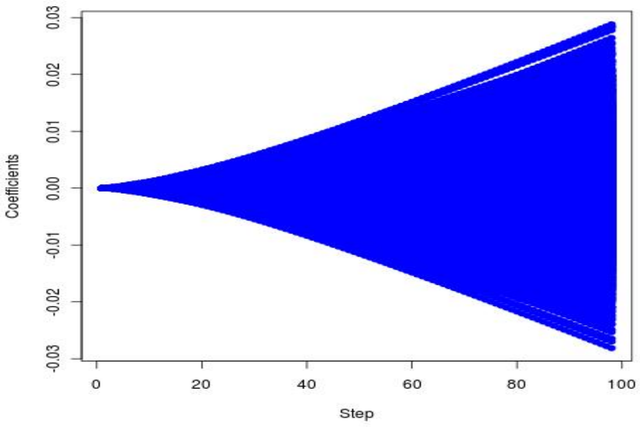

Figure 4: Profile of estimated coefficients for GLMNET method

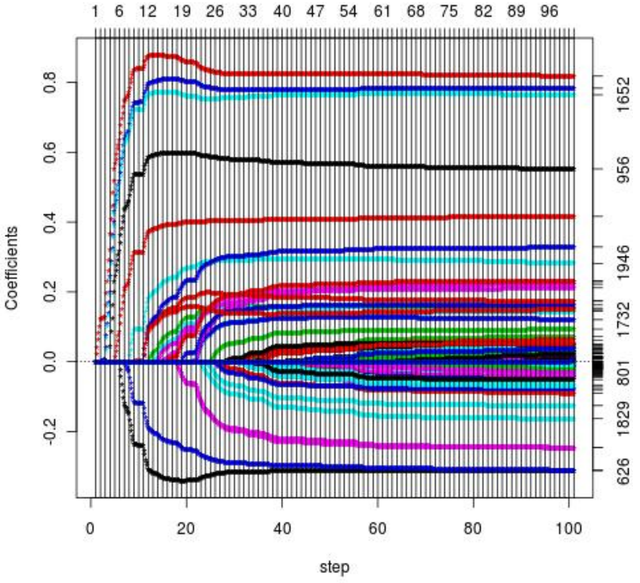

Figure 5: Profile of estimated coefficients for CGGD method

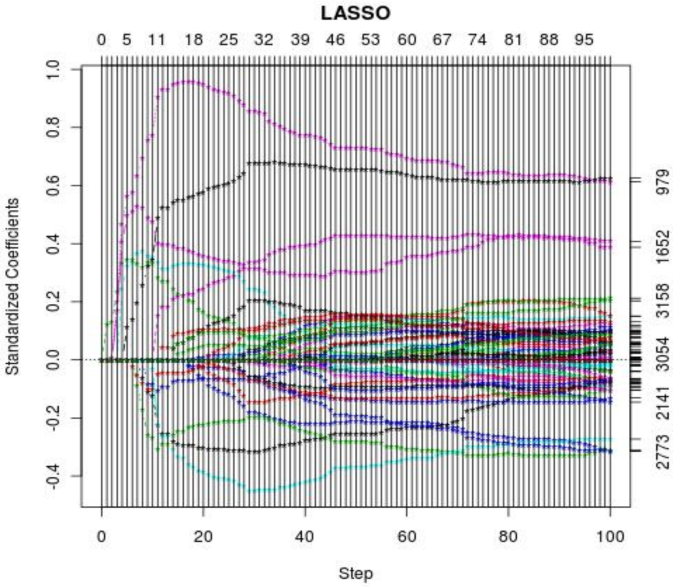

Figure 6: Profile of estimated coefficients for LARS method

10 Fold cross validation was used, and timings were recorded. Timing in seconds for GLMNET, CGGD, Relaxo, and LARS over Leukemia data is presented. The timings were performed on one HP TX2000 series laptop.

Optimization 100 Steps	
GLMNET	0.036s
Relaxo	0.064s
LARS	0.116s
CGGD	1.280s

Cross-validation 100 Steps	
GLMNET	0.420s
CGGD	1.38s
LARS	1.932s
Relaxo	4.076s

Conclusions

When compared, GLMNET is the more efficient algorithm. By the 100th step the predicted coefficients for GLMNET are stronger than both CGGD and LARS. When comparing the timings, GLMNET is almost 4 times as quick as CGGD in both optimization and cross validation. Relaxo is the almost twice as slow as GLMNET when comparing optimization and almost 10 times as slow when cross validating. We can conclude that the most efficient method for L1-regularized logistic regression is GLMNET. The Leukemia dataset has a larger number of features compare to the number of instances. Linear models work well with datasets with such characteristics. The data while large however contained a small number of samples. Testing over a dataset with a large sample and small feature should be further investigated.

References

Dettling M. 2004. BagBoosting for Tumor Classification with Gene Expression Data. *Bioinformatics*, 20, 3583-3593.

Efron, B.; Hastie, T.; Johnstone, I.; and Tibshirani, R. 2004. Least angle regression. *Annals of Statistics,* 32:407-499.

Friedman, J.; Hastie, T.; and Tibshirani, R. 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* 33(1):1-22

Golub T.; Slonim D.K.; Tamayo P.; Huard C.; Gaasenbeek M.; Mesirov J.P.; Coller H.; Loh M.L.; Downing J.R.; Caligiuri M.A.; Bloomfield C.D.; Lander E.S. 1999. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, 286, 531-536.

Hastie, T.; Rosset, S.; Tibshirani, R.; and Zhu, J. 2004. The entire regularization path for the support vector machine. *Journal of Machine Learning Research*, *5*:1391–1415.

Koh, K.; Kim, S. J.; and Boyd, M. 2007. An Interior-Point Method for Large-Scale L1-Regularized Logistic Regression. *Journal of Machine Learning Research* 8:1519-1555.

Kuhn M. 2012. Package ‘caret’ http://cran.r-project.org/web/packages/caret/caret.pdf

Meinshausen N. 2007. Relaxed Lasso. *Computational Statistics and Data Analysis* 52(1), 374-393

Wainwright, M.; Ravikumar, P.; and Lafferty, J. 2007. High-dimensional graphical model selection using L1-regularized logistic regressionn. *Advances in Neural Information Processing Systems (NIPS) 19*.

Yuan, G. X.; Chang, K. W.; and Lin, C. J. 2010. A Comparison of Optimization Methods and Software for Large-scale L1-regularized Linear Classification. *Journal of Machine Learning Research* 11: 3183-3234.

Yuan, G. X.; Ho, C. H.; and Lin, C. J. 2011. . An Improved GLMNET for L1-regularized Logistic Regression. *In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Zhang, C. H. 2007. Continuous Generalized Gradient Descent. *Journal of Computational and Graphical Statistics.*